

**PENERAPAN ALGORITMA BIDIRECTIONAL ENCODER
FROM TRANSFORMER UNTUK MENDETEKSI
PLAGIARISME PADA KODE
PROGRAM BAHASA C++**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

BAYU TAUFIQURRAHMAN

18.11.2086

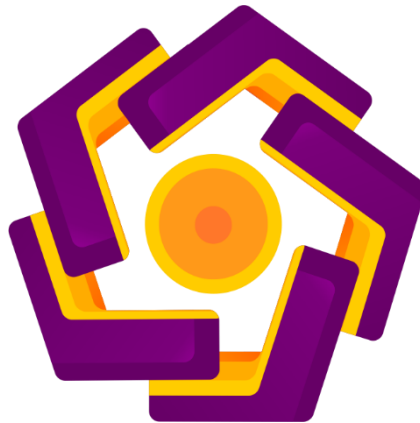
Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023**

**PENERAPAN ALGORITMA BIDIRECTIONAL ENCODER
FROM TRANSFORMER UNTUK MENDETEKSI
PLAGIARISME PADA KODE
PROGRAM BAHASA C++**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

BAYU TUFIQURAHMAN

18.11.2086

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2023

HALAMAN PERSETUJUAN

HALAMAN PERSETUJUAN

SKRIPSI

**PENERAPAN ALGORITMA BIDIRECTIONAL ENCODER FROM
TRANSFORMER UNTUK MENDETEKSI PLAGIARISME
PADA KODE PROGRAM BAHASA C++**

yang disusun dan diajukan oleh

Bayu Taufiqurrahman
18.11.2086

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 10 Juli 2023

Dosen Pembimbing,



Yoga Pristyanto, S.Kom, M.Eng
NIK. 190302412

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

SKRIPSI

PENERAPAN ALGORITMA BIDIRECTIONAL ENCODER FROM TRANSFORMER UNTUK MENDETEKSI PLAGIARISME PADA KODE PROGRAM BAHASA C++

yang disusun dan diajukan oleh

Bayu Taufiqurrahman

18.11.2086

Telah dipertahankan di depan Dewan Penguji
pada tanggal 31 Juli 2023

Susunan Dewan Penguji

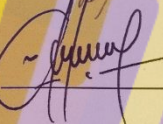
Nama Penguji

Yoga Pristvanto, S.Kom., M.Eng
NIK. 190302412

Dr. Ferry Wahyu Wibowo, S.Si., M.Cs.
NIK. 190302235

Majid Rahardi, S.Kom., M.Eng
NIK. 190302393

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 31 Juli 2023

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta, S.Kom., M.Kom.
NIK. 190302096



HALAMAN PERNYATAAN KEASLIAN SKRIPSI

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Bayu Taufiqurrahman
NIM : 18.11.2086

Menyatakan bahwa Skripsi dengan judul berikut:

Penerapan Algoritma Bidirectional Encoder From Transformer Untuk Mendeteksi Plagiarisme Pada Kode Program Bahasa C++

Dosen Pembimbing : Yoga Pristyanto, S.Kom, M.Eng

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 31 Juli 2023

Yang Menyatakan,



Bayu Taufiqurrahman

KATA PENGANTAR

Assalamu'alaikum warahmatullahi wabarakatuh.

Puji Syukur kehadiran Allah SWT atas limpahan rahmat dan hidayah-Nya, penulis dapat menyelesaikan tugas akhir ini dengan judul “Penerapan Algoritma Bidirectional Encoder From Transformer Untuk Mendeteksi Plagiarisme Pada Kode Program Bahasa C++”. tugas akhir ini merupakan salah satu persyaratan yang harus peneliti penuhi untuk memperoleh gelar sarjana Fakultas Ilmu Komputer Jurusan Informatika di Universitas Amikom Yogyakarta. Selama proses penulisan tugas akhir ini, penulis menghadapi berbagai hambatan, namun berkat dukungan, bimbingan, dan kerjasama yang tulus dari berbagai individu, akhirnya tugas akhir ini berhasil diselesaikan dengan sukses. penulis mengucapkan terima kasih kepada:

1. Bapak Yoga Pristyanto, S.Kom, M.Eng selaku dosen pembimbing yang telah memberikan bimbingan, arahan, motivasi serta kesabaran kepada penulis selama pengerjaan tugas akhir sehingga tugas akhir ini dapat terselesaikan.
2. Bapak Dr. Ferry Wahyu Wibowo, S.Si., M.Cs., Bapak Majid Rahardi, S.Kom., M.Eng selaku dosen penguji yang telah memberikan saran untuk perbaikan tugas akhir ini.
3. Kedua orang tua serta wali penulis yang selalu memberikan doa, motivasi dan dukungan moral dalam menyelesaikan tugas akhir ini.
4. Seluruh jajaran dosen dan staff Universitas Amikom Yogyakarta.
5. Teman-teman IF04 dan rekan kerja student staff DAAK Pengajaran yang saling mendukung.

Penulis memahami bahwa penelitian dalam tugas akhir ini masih jauh dari kesempurnaan, oleh karena itu, penulis mengundang pembaca untuk memberikan kritik dan saran. Harapan penulis adalah agar hasil dari tugas akhir ini dapat memberikan nilai dan manfaat yang signifikan bagi semua pihak yang terlibat.

Yogyakarta, 31 Juli 2023

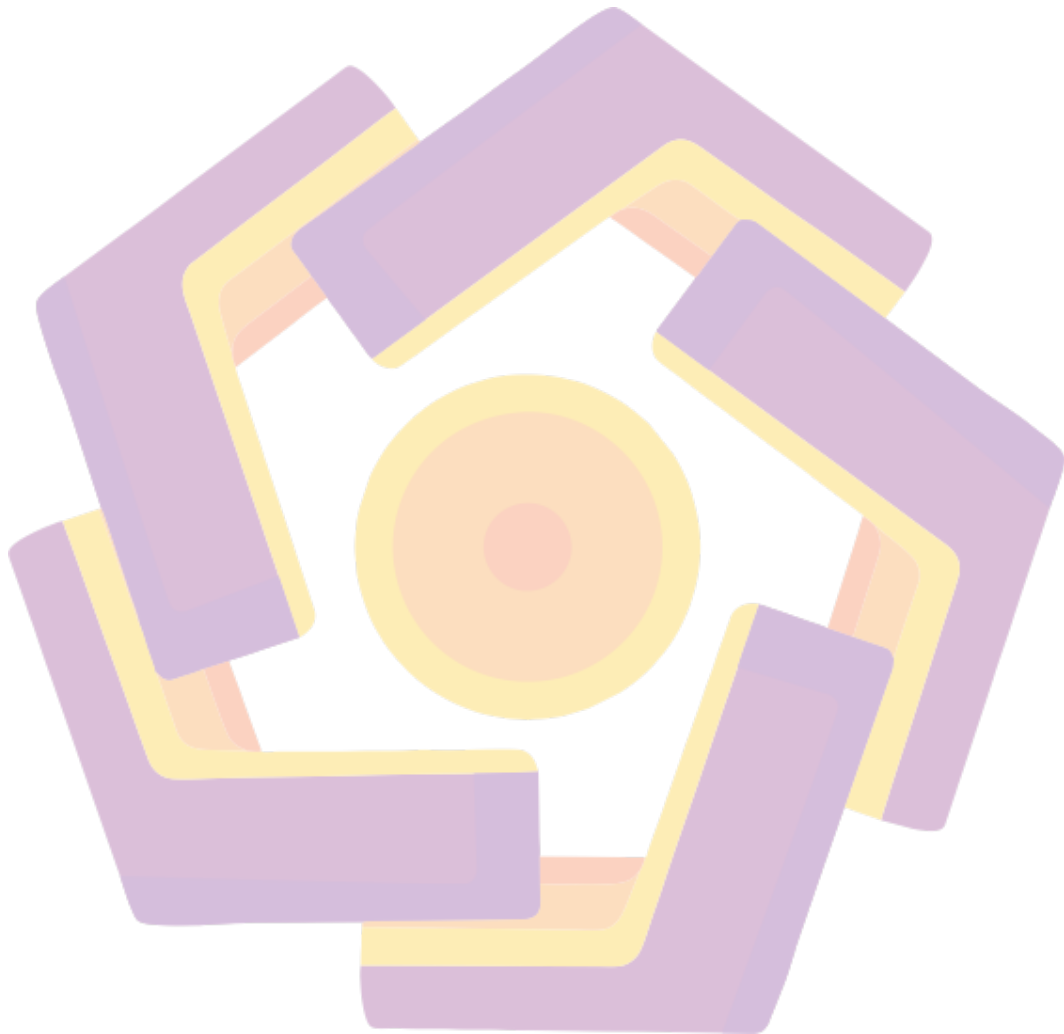
Penulis

DAFTAR ISI

PENERAPAN ALGORITMA BIDIRECTIONAL ENCODER FROM TRANSFORMER UNTUK MENDETEKSI PLAGIARISME PADA KODE	i
PROGRAM BAHASA C++.....	i
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
KATA PENGANTAR	v
DAFTAR ISI	vi
DAFTAR TABEL	ix
DAFTAR GAMBAR	x
INTISARI.....	xi
ABSTRACT	xii
BAB 1	1
PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	2
1.4 Maksud dan Tujuan Penelitian	2
1.5 Manfaat Penelitian	3
2.2 Sistematika Penulisan	3
BAB II.....	4
LANDASAN TEORI.....	4
2.1 Tinjauan Pustaka.....	4
2.2 Dasar Teori.....	10
2.2.1 <i>Natural Language Processing</i>	10
2.2.2 <i>Deep Learning</i>	11
2.2.3 Plagiarisme	11
2.2.4 C++	12

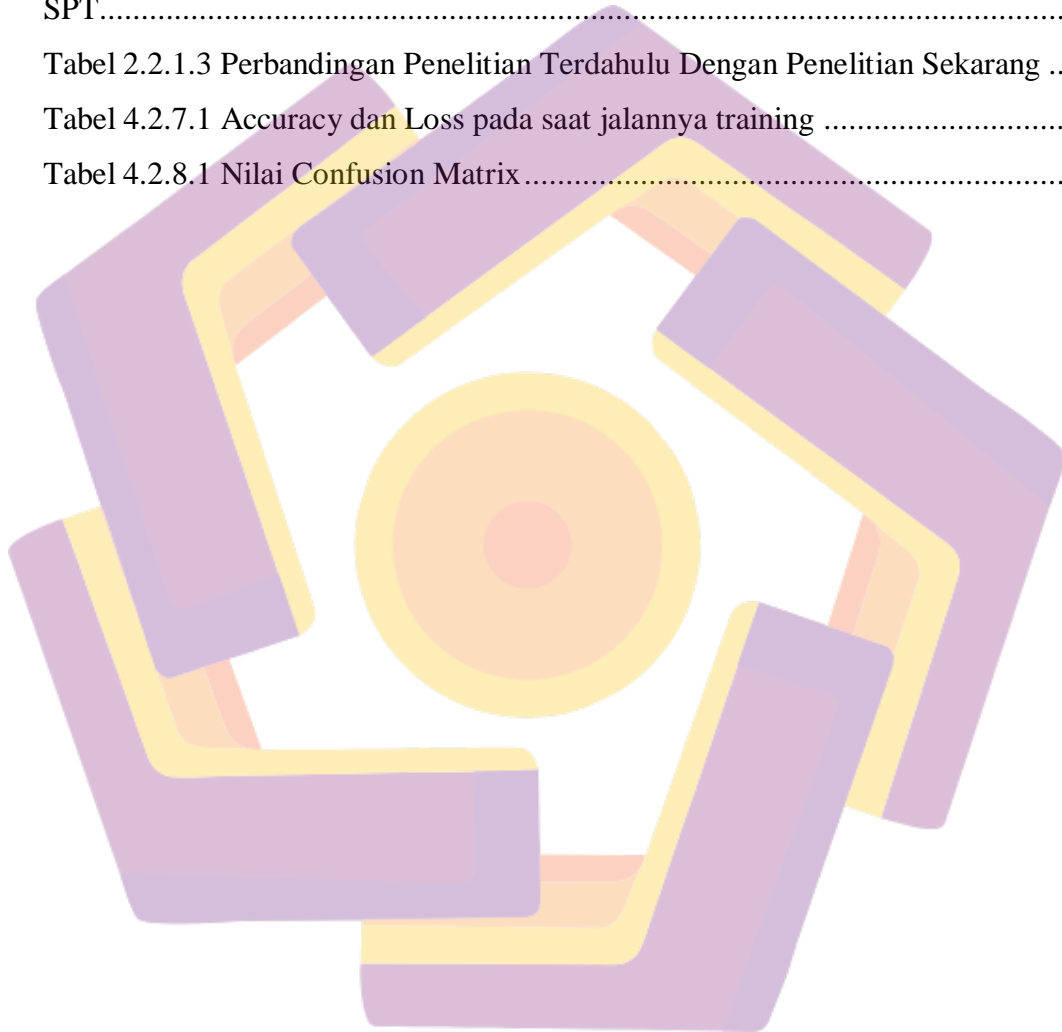
2.2.5	<i>Bidirectional Encoder Representations From Transformers (BERT)</i>	
		12
2.2.6	<i>Confusion Matrix</i>	15
BAB III	17
METODE PENELITIAN	17
3.1	Alat dan Bahan	17
3.1.1	Alat Penelitian	17
3.1.2	Bahan Penelitian	18
3.2	Alur Penelitian.....	18
3.2.1	Pengumpulan Data.....	19
3.2.3	Persiapan Data.....	19
3.2.3.1	Convert Dataset.....	19
3.2.3.2	Preprocessing Data.....	19
3.2.3.3	Data Visualization.....	19
3.2.3.4	Data Tokenization.....	20
3.2.3.5	Data Split.....	20
3.2.4	<i>BERT Model</i>	20
3.2.5	<i>Fine Tuning</i>	20
3.2.6	<i>Model Evaluation</i>	20
BAB IV	20
HASIL & PEMBAHASAN	20
4.1	Pengumpulan Data.....	20
4.2	Persiapan Data.....	20
4.2.1	<i>Convert Dataset</i>	20
4.2.2	<i>Preprocessing Data</i>	21
4.2.3	<i>Data Visualization</i>	23
4.2.4	<i>Data Tokenization</i>	24
4.2.5	<i>Data Split</i>	24
4.2.6	<i>BERT Model</i>	25
4.2.7	<i>Fine Tuning</i>	27
4.2.8	<i>Model Evaluation</i>	30

BAB V.....	36
KESIMPULAN & SARAN	36
5.1 Kesimpulan.....	36
5.2 Saran	36
DAFTAR PUSTAKA	37



DAFTAR TABEL

Tabel 2.2.1.1 Akurasi persentase MLP with Bag of Tokens dan Siamese Network with Token Sequence	6
Tabel 2.2.1.2 Scoring SPT with Handcrafted Feature Extraction dan GNN with SPT.....	6
Tabel 2.2.1.3 Perbandingan Penelitian Terdahulu Dengan Penelitian Sekarang	7
Tabel 4.2.7.1 Accuracy dan Loss pada saat jalannya training	28
Tabel 4.2.8.1 Nilai Confusion Matrix	35



DAFTAR GAMBAR

Gambar 2.2.1.1 Paste Feature.....	4
Gambar 2.2.2.1 Ilustrasi deep learning menggunakan dataset mnist handwritten digit	11
Gambar 2.2.3.1 Tabel presentase tingkat plagiarisme	12
Gambar 2.2.5.1 Bert Layer Encoder	13
Gambar 2.2.5.2 Model Bert.....	13
Gambar 2.2.6.1 Confusion Matrix	15
Gambar 3.2.6.1 Kode Pemrograman C++ Perkalian Perulangan	20
Gambar 4.2.1.1 Kategori Parsing	20
Gambar 4.2.1.2 Potongan path file teks	21
Gambar 4.2.1.3 Potongan proses parsing	21
Gambar 4.2.2.1 Output Preprocessing tahap ke-1	21
Gambar 4.2.2.2 Info dataset setelah dikurangi	22
Gambar 4.2.2.3 Info dataset setelah merging	22
Gambar 4.2.2.4 Label Encoding	23
Gambar 4.2.2.5 Dataset setelah melalui tahap preprocessing	23
Gambar 4.2.3.1 Avg Len.....	23
Gambar 4.2.4.1 Proses Tokenisasi BERT	24
Gambar 4.2.5.1 Output data setelah proses split	25
Gambar 4.2.6.2 Output bert base cased model.....	25
Gambar 4.2.7.1 Full Model	27
Gambar 4.2.7.2 Grafik Loss	29
Gambar 4.2.7.3 Grafik Accuracy.....	29
Gambar 4.2.8.1 Confusion Matrix Plagiarism C++	30

INTISARI

Plagiarisme merupakan sebuah tindakan yang mengambil karya, ide, ataupun informasi tanpa sepengetahuan pemilik hak cipta kemudian mengakuinya atas miliknya sendiri. Tugas akhir pemrograman setiap semester rentan terhadap plagiarisme oleh mahasiswa. Untuk mendeteksi adanya perbuatan plagiarisme, penelitian ini menggunakan algoritma bidirectional encoder from transformer. Algoritma bidirectional encoder from transformer diperkenalkan oleh Google untuk meningkatkan ketepatan mesin pencarian miliknya.

Penelitian ini dilakukan melalui beberapa tahap yaitu (1) data cleansing tujuannya untuk membuang data yang tidak dapat terpakai agar tidak menyebabkan overfitting maupun underfitting. (2) Visualisasi data merupakan tahap mengetahui data yang digunakan untuk training. (3) Training dan testing data merupakan sebuah proses untuk melatih dan mengetahui hasil dari algoritma yang diterapkan.

Dataset yang digunakan untuk penelitian ini berasal dari IBM dengan total seribu empat ratus kode program bahasa C++ yang setiap kodenya sudah dilabeli dengan waktu proses cpu, dan penggunaan memori yang digunakan saat program dijalankan. Dataset ini akan dilakukan preprocessing terlebih dahulu seperti data cleansing dan tokenizing. Hasil dari penelitian diharapkan dapat membuat para pengajar lebih mudah untuk mengatasi plagiarisme kode program bahasa C++.

Kata Kunci: bert, nlp, transformer, machine learning, plagiarisme

ABSTRACT

Plagiarism is an act of taking works, ideas, or information without the knowledge of the copyright owner and then acknowledging it as his own. Final project on programming per semester is easily plagiarized by students. To detect plagiarism, this research uses a bidirectional encoder from transformer algorithm. The bidirectional encoder from transformer algorithm was introduced by Google to improve the accuracy of its search engine.

This research was carried out through several stages, namely (1) data cleansing, which was to dispose of data that could not be used so as not to cause overfitting or underfitting. (2) Data visualization is the stage of knowing the data used for training. (3) Training and data testing is a process to train and find out the results of the applied algorithm.

The dataset used for this research comes from IBM with a total of one thousand four hundred C++ language program codes, each of which has been labeled with cpu processing time, and the memory usage used when the program is run. This dataset will be preprocessed first, such as data cleansing and tokenizing. The results of the research are expected to make it easier for teachers to overcome plagiarism in the C++ language programming.

Keyword: bert, nlp, transformer, machine learning, plagiarism